DOI:
10.1111/cgf.13169

Document Version
Peer reviewed version

Eurographics Conference on Visualization (EuroVis) 2017
J. Heer, T. Ropinski and J. van Wijk
(Guest Editors)

Volume 36 (2017), Number 3

Empirically Measuring Soft Knowledge in Visualization

Natchaya Kijmongkolchai, Alfie Abdul-Rahman, and Min Chen

University of Oxford, UK

Abstract

*In this paper, we present an empirical study designed to evaluate the hypothesis that humans' soft knowledge can enhance the cost-benefit ratio of a visualization process by reducing the **potential distortion**. In particular, we focused on the impact of three classes of soft knowledge: (i) knowledge about application contexts, (ii) knowledge about the patterns to be observed (i.e., in relation to visualization task), and (iii) knowledge about statistical measures. We mapped these classes into three control variables, and used real-world time series data to construct stimuli. The results of the study confirmed the positive contribution of each class of knowledge towards the reduction of the potential distortion, while the knowledge about the patterns prevents distortion more effectively than the other two classes.*

1. Introduction

The knowledge of the viewers is an intrinsic part of a visualization process. It can enable selective attention, pattern recognition, and visual reasoning, but at the same time it may cause inattentional blindness, illusion, and biases. Although this has been widely accepted intuitively, it has been a challenge to measure the effect of such knowledge in visualization. Most laboratory-based empirical studies in the literature focused on the effects of different visual mappings, often in conjunction with an effort to minimize the confounding impact of the variance among participants' knowledge.

Recently, Chen and Golan outlined an information-theoretic explanation of humans' soft knowledge in visualization processes [CG16]. As illustrated in Figure 1, consider all possible data objects (e.g., all possible time series) as a data alphabet $\mathbb{D}$, and all possible visual objects of a particular visual representation (e.g., a polyline in a display) as a visualization alphabet $\mathbb{V}$. The visual encoding of a data object $d \in \mathbb{D}$ yields a visual object $v \in \mathbb{V}$. When a human viewer observes the visual object v, he/she may have some knowledge about the data alphabet, e.g., the practical validity and the expected likelihood of letters in $\mathbb{D}$. For example, a self-intersected polyline does not represent a valid time series of temperature. A highly oscillated polyline is an unusual time series for a stock price.

Using the term defined in [CG16], soft knowledge can facilitate the reduction of *potential distortion* in decoding, i.e., the reverse mapping from $\mathbb{V}$ to $\mathbb{D}$. Information-theoretically, as illustrated in Figure 1, the soft knowledge provides the data alphabet $\mathbb{D}$ with additional constraints on its probability mass function. Since soft knowledge is sensitive to the variation of applications, users, and visualization tasks, the changes to the probability mass functions usually differ from one context to another, and often vary dynamically within an interactive visualization process.

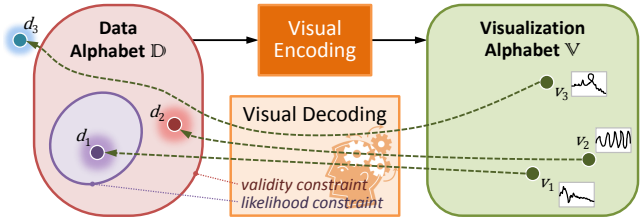


Figure 1: *During visualization, human viewers are able to judge if a visual object is valid (e.g., v_1), unlikely (e.g., v_2), or invalid (e.g., v_3) using their soft knowledge about the probability distribution of the data alphabet. The halos around d_1, d_2, and d_3 indicate that visual decoding features uncertainty since visual encoding is usually a many-to-one mapping.*

While this information-theoretic explanation indicates the benefit of visualization processes, it has not yet been confirmed empirically by any purposely-defined study. In [CG16], there was no suggestion as to how to measure the capacity of different soft knowledge. In [TKC17], an attempt was made to estimate soft knowledge analytically. Because the estimated quantities were derived based on the observations of two expert analysts who performed highly complex model development tasks, it would be difficult to confirm the estimated quantities using a laboratory-based empirical study.

At the beginning of this work, we regarded the information-theoretic explanation about the role of soft knowledge in visualization [CG16] as an unconfirmed hypothesis. Our objective is to design and conduct an empirical study to evaluate this hypothesis. In particular, we constructed our stimuli using real-world time series data and their corresponding time series plots. We examine the impact of three classes of soft knowledge: (i) knowledge about application contexts, (ii) knowledge about the patterns to be observed

Table 1: *A categorization of variables examined in various empirical studies reported in the visualization literature.*

Perception & Cognition	**Context**	**Pattern**	**Statistics**
Memory in visualization	[BBK*16]	[BVB*13]	
Attraction & enjoyment		[SSK16] [DBD17]	
Attention in visualization		[HW12]	
Factors affecting visual reasoning	[MDF12] [LARC16] [OPH*16]	[MDF12] [LARC16]	[OPH*16]
Visual grouping	[BW14]	[BNRS13] [GSL14] [BW14]	[GSL14]
Laws & models		[HYFC14]	[HYFC14] [TGH12]
Colors	[LFK*13] [MK15]	[BPC*10]	[BPC*10]
Shapes		[GHL15]	[GHL15]
Size		[GSL14]	[GSL14] [JH16]
Numbers		[BDJ14]	[GCNF13] [BDJ14]
Order			[CAB*16]
Dimension coverage	[STM17]		
Spatial autocorrelation		[BDM*17]	
Effects of display attributes	[JH13]	[BI12]	
Effects of layout features		[MPWG12] [ZOC*13]	
Scatter plots	[VTW*12]	[LMVW10] [VTW*12] [FHSW13] [KARC15]	[LMVW10] [RB10] [KARC15]
Perception of parallel coordinates		[LMVW10] [KZZM12] [KARC15]	[LMVW10] [KZZM12] [KARC15]
Trust in visualization	[DLW*17]		
Effects of visual embellishments	[BARM*12]	[BARM*12] [SHK15]	[BARM*12] [SHK15]

Comparison of Techniques	**Context**	**Pattern**	**Statistics**
Flow visualization	[BBL12] [YDGM17]	[LDM*01] [BBL12] [YDGM17]	[YDGM17]
Volume visualization		[ZWM13]	
Graph visualization	[WCA*17]	[MPWG12] [XRP*12] [FIB*14] [SSKB14] [SSK16] [KMLM16]	
Tree visualization		[ZOC*13]	
Time-series data	[WBJ16]	[ARH12] [HKF16] [WBJ16]	
Text visualization		[SOK*16]	
Geo-visualization		[SSKB15] [GR15] [NHB*17] [BDM*17]	
Multivariate data visualization	[EMdSP*15]	[LDA12] [EMdSP*15]	[EMdSP*15]
Trajectory visualization		[NBW14]	
Pixel-based visualization		[BPC*10] [PQMCR17]	[BPC*10] [PQMCR17]
Interactivity	[JHKH13] [BEDF16]	[RKSB13] [KDX*12] [BEDF16]	
Animation		[VBC*16] [CDF14]	
Video visualization	[HKH*12]	[CBH*06] [HKH*12] [KHW13]	
Statistical charts & plots		[TSA14] [SK16] [ALBR16]	[TSA14] [SK16] [ALBR16]
Different coordinate systems		[HFMC12]	
Visual encodings of uncertainty		[BBIF12] [CG14] [MRO*12] [GBFM16]	[CG14]
Highlighting techniques		[GR15] [SOK*16]	
Visualization authoring		[WPHC16]	
Online visualization education	[TLM16]	[TLM16]	

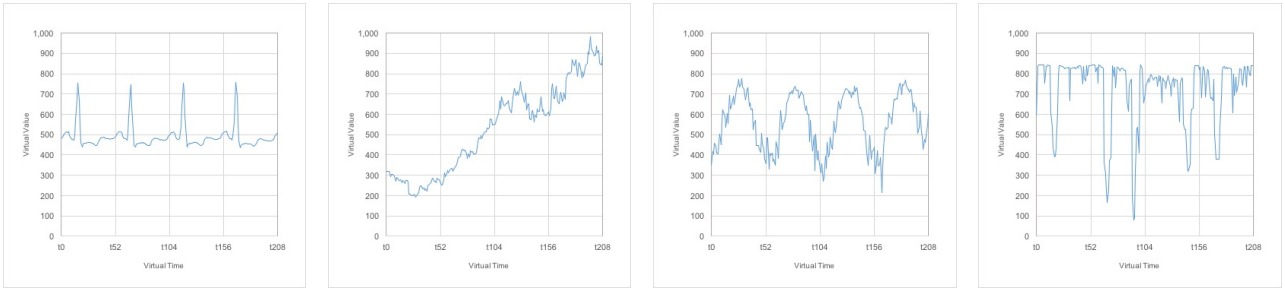
(i.e., in relation to visualization task), and (iii) knowledge about the mathematical definitions of statistical measures. Our study confirmed that all three classes of knowledge have positive impact on human participants' inference. When the probabilistic distribution of the data space is equally affected by each class of knowledge, participants' knowledge about the patterns had a higher impact on the participants' performance than the other two classes.

2. Related Work

Paul Cohen described four types of empirical studies: *Exploratory Studies*, *Assessment Studies*, *Manipulation Experiments*, and *Observation Experiments* [Coh95] (p.7). The recent theoretical work by Chen and Golan [CG16] includes a number of case studies, which can collectively be considered as an exploratory study that has resulted in a hypothesis of humans' soft knowledge can reduce *potential distortion* in data intelligence processes. As stated in [Coh95], hypotheses established in exploratory studies are normally tested in observation or manipulation experiments. The recent study by Tam et al. [TKC17] is an observation experiment that made naturalistic observations of two analysts' contributions to the construction of different classification models. They narrowed down the soft knowledge to a number of variables referred to as "observations", and used information theory to estimate the quantities of such knowledge. However, it is highly desirable to conduct a manipulation experiment that can test the hypothesis in [CG16] by manipulating some factors about knowledge and "noting the effects, if any, on one or more measured variables." [Coh95].

In the field of visualization, there are many empirical studies. Many manipulation experiments were independently reported in the literature, while most observation studies were usually reported in conjunction with design studies and applications. As shown in Table 1, there are many experiments on the perception and cognition in visualization and many for comparing different methods. Although the table does not include all empirical studies in visualization, it provides an extensive coverage of the typical topics being studied. To the best of our knowledge, there has not been a manipulation experiment focusing explicitly on soft knowledge in visualization. However, many studies feature dependent variables that represent different levels or types of soft knowledge. Here we coarsely categorize such variables into three categories:

- **Context** — This encompasses variables that change the underlying data spaces, e.g., different applications. We also include variables about algorithms and interactions that result in changes to participants' attention to different parts of a data space.



(a) ECG (2.496 seconds) (b) Stock (4 years) (c) Temperature (4 years) (d) Electricity (4 years)

Figure 2: *Time series captured in different contexts usually have some characteristic visual signatures. Through their lives, humans have acquired, often unconsciously, the knowledge for recognizing the visual signatures of commonly-seen time series.*

- **Pattern** — This encompasses variables that change the visual representations of the same data. They are typically used for examining participants' performance in observing relatively complex information (e.g., features, patterns, events, etc.) rather than individual numbers.
- **Statistics** — This encompasses variables that change the visual representations related to specific numbers and statistical measures. They are typically used for examining participants' performance in determining the values of individual measures.

In this work, instead of examining these variables from the perspective of visual encoding, we investigate whether such variables may indicate the presence of humans' soft knowledge in visualization and their impact on humans' inference. In particular, it is designed to evaluate the hypothesis that such soft knowledge brings tangible benefits to inference because visualization enables human observers to use their knowledge. We believe that this is the first manipulation experiment that focuses directly on humans' soft knowledge in visualization. It is also the first attempt to measure such benefits through an empirical study.

3. Methodology

In [CG16], Chen and Golan proposed an information-theoretic metric for measuring the cost-benefit ratio of data analysis and visualization workflows as well as their components:

$$\frac{\text{Benefit}}{\text{Cost}} = \frac{\text{Alphabet Compression} - \text{Potential Distortion}}{\text{Cost}} \quad (1)$$

Traditionally, the "goodness" of a machine- or human-centric process is measured by the accuracy of its output. However, in many situations, the correctness of possible outputs (e.g., financial decisions) is not well-defined. The above cost-benefit metric avoids the subjectiveness in measuring the correctness of an output. Instead, it measures the balance between its ability to abstract (i.e., *Alphabet Compression*) and its ability to reconstruct (i.e., less *Potential Distortion*). It suggests that humans' soft knowledge in visualization processes can reduce the potential distortion.

For example, consider an alphabet of a time series as illustrated in Figure 1. The probability of each time series (i.e., letter) in the alphabet is sensitive to many factors. One important factor is the application context. Figure 2 shows four time series obtained from real-word data repositories in four different contexts, namely electrocardiogram (ECG), stock market, weather temperature, and electricity production. Without viewing any other data in such repositories, most human observers would consider that the patterns of the time series in Figure 2(a) is less probable in a data repository of stock market or weather temperature. Through their lives, humans have acquired, often unconsciously, the knowledge for recognizing the visual signatures of commonly-seen time series. In addition to the four contextual types in Figure 2, experts and non-experts may also have accumulated knowledge about rainfall record, tidal height, seismograph, electroencephalogram (EEG), photoplethysmogram (PPG), exchange rate, gross domestic product (GDP), electricity consumption, and so on. In abstract, the ability to recognize the visual signatures of time series in a context is more or less the same as the ability to have the intuition (i.e., tacit knowledge) about the probability distribution of an alphabet in the context.

In an application context, there are usually many specific categorical terms describing various temporal patterns in time series. For example, in the context of ECG, *atrial flutter* is an abnormal heart condition, which produces a very distinct sawtooth pattern in an ECG signal. Meanwhile, *ventricular fibrillation* is a serious abnormal heart condition, which produces a highly irregular pattern, very different from a normal ECG. Cardiologists usually recognize such patterns at a glance. Such a sophisticated decision process can also be translated to reliable "reconstruction" in knowing the probability of different features and patterns of time series in the ECG alphabet in relation to a diagnostic decision.

In general, as illustrated in Figure 1, visualization enables human observers to use their soft knowledge to establish a coarse conditional probability distribution in a certain context, under a certain condition, for a certain task requirement, or with certain information. Although such soft knowledge seems rather effective in practice, the "gut feeling" about the conditional probability distribution also seems rather imprecise numerically. Although the cost-benefit metric in [CG16] can explain the benefit of soft knowledge, one would still need some scientifically verifiable evidence to support this. This empirical study was designed to examine the effectiveness of such gut feeling about a conditional probability distribution.

The analytical evidence in [TKC17] shows that humans possess a tremendous amount of knowledge that could be deployed in a visualization process. In order to measure such knowledge in a statistically meaningful way, an empirical study has to focus on some particular pieces of knowledge. In this study, we measure human

Table 2: *Textual and binary encoding of eight optional answers.*

Textual	Binary	Type	Pattern	Statistics
TPS	111	matching	matching	matching
TPs	110	matching	matching	mismatched
TpS	101	matching	mismatched	matching
Tps	100	matching	mismatched	mismatched
tPS	011	mismatched	matching	matching
tPs	010	mismatched	matching	mismatched
tpS	001	mismatched	mismatched	matching
tps	000	mismatched	mismatched	mismatched

knowledge in the form of three *soft models* [TKC17]. The first model, $\mathbf{M}_T$, is a function that takes the input of a time series alphabet $\mathbb{A}$ and a clue C_T about the contextual type of the application, and divides the alphabet into two groups, one matches C_T, and one does not. The second model, $\mathbf{M}_P$, takes the input of $\mathbb{A}$, and a clue C_P about a specific pattern to be observed, and divides the alphabet into two groups, one matches C_P, and one does not. The third model, $\mathbf{M}_S$, takes the input of $\mathbb{A}$, and a clue C_S about a statistical measure as a requirement, and divides the alphabet into two groups, one matches C_S, and one does not.

As shown in Table 2, we can assign each letter in $\mathbb{A}$ a textual or binary code according to whether it matches with the three clues C_T, C_P, and C_S. When alphabet $\mathbb{A}$ has exactly one letter with each of the eight codewords, the combined use of the three soft models should ideally derive the letter coded as (TPS, 111). When the inference process derives a letter with a code other than (TPS, 111), there is a failure by at least one soft model. For example, if the resultant letter is coded as (TpS, 101), it is the failure of $\mathbf{M}_P$; or if (tps, 000), it is the failure of all three soft models $\mathbf{M}_T$, $\mathbf{M}_P$, and $\mathbf{M}_S$. When we repeat this exercise for different alphabets, we can obtain quantitative measures about the success rate of each model.

4. Study Design

Trial Design and Task Design. Each trial tests the usage of three pieces of knowledge $\mathbf{M}_T$, $\mathbf{M}_P$, and $\mathbf{M}_S$ in selecting one time series plot from 8 optional answers. The three pieces of knowledge were stimulated in two stages. Firstly, a participant is shown an *information screen*, an example of which is shown in Figure 3. All information screens are designed in the form of a magazine or newspaper article. Among the text and image(s) on each information screen, there are two clues that relate to a contextual type (e.g., ECG, Stock, or Temperature), and a specific pattern (e.g., up, down, flat, etc.) respectively. The two clues in Figure 3 are "ECG" and "wandering baseline trending upwards". The clues are designed to be easy to remember. In a real-world scenario, a person may use such clues unconsciously without any explicit prompt, since they are usually defined by the role of, and the task performed by, the person (e.g., a doctor inspecting a patient record). Because the participants had to experience 36 different scenarios, a prompt is unavoidable. Nevertheless, the clues on the information screens are designed to be vague but easy to remember, and will not be available when the participants are performing their tasks.

There is no temporal restriction as to how long a participant can read the information screen. Once a participant presses the "next" button, the trial moves to the second stage. A corresponding *question screen* replaces the information screen. Figure 4 shows the question screen that follows the information screen in Figure 3. On the top of the screen, there is a third clue describing the statistical indicator. Because such information is difficult to remember and in a real-world scenario, a person normally acquires such information consciously for a task, we therefore made sure that this clue was always available when a participant was selecting an answer from the optional time series plots. In the question screen, the 8 optional plots are randomly placed in a 4×2 grid. In Figure 4, C is the correct answer that matches all three clues (TPS, 111), and the 7 distractors are: A: (tpS, 001), B: (tPS, 011), D: (tPs, 010), E: (TPs, 110), F: (tps, 000), G: (Tps, 100), and H (TpS, 101). All these distractors were carefully designed to avoid giving out any further hint in addition to the mismatches with the given clues. For example, all 4 type-distractors in Figure 4 show temperature data in order to avoid a hint due to unequal numbers of contextual types.

Another three examples of information screens are shown in Figure 5. Figure 5(a) is in the form of a newspaper article. It introduces the concept of "bear market" to participants, giving clues about a contextual type of stock market and a specific pattern of downward trending. The other two are in the form of magazine articles. Figure 5(b) describes the weather of Thailand, giving clues about a contextual type of weather temperature and a specific pattern of high temperatures in April and May, which are to be interpreted as a visual signature of a peak around 1/3 length into each of the four temporal sections. Figure 5(c) describes seasonal patterns of wind energy, giving clues about a contextual type of electricity production and a specific seasonal pattern in New England. All information screens and question screens used in this study are included in the supplementary materials.

In each trial, the task of each participant is to select a time series plot that matches all three clues. There are a total of 36 trials, each has the same composition of 8 optional answers as defined in Table 2 and exemplified in Figure 4. The number of 36 is determined by 6 type combinations × 2 classes of patterns × 3 statistical measures.

There are three contextual types, *electrocardiogram (ECG)*, *stock price*, and *weather temperature*. Because each trial features two contextual types, one correct type and one distractor type, there are thus six type-combinations.

In terms of specific patterns, we made use of patterns in the real-world data. Therefore, patterns vary noticeably from one contextual type to another. We divided patterns roughly into two groups, *Simple* and *Complex*. The simple patterns require little semantic interpretation, such as signals featuring upward and downward trending, long flat or near flat segments, normal or plain patterns, and so on. For example, in terms of stock prices, there are "slowly trending up", "slowly trending down", "sharp rise", "sharp drop", "stable", and "missing a section of data". The complex patterns usually appear in specific parts of a time series plot, requiring some semantic reasoning and effort to search for a pattern across the time series. For example, in terms of stock prices, there are "January effect", "Superbowl impact", "outlier values", "bull trap", "high volatility", and "anomalous calm".

For statistical indicators, we consider three commonly used measures, *minimum or maximum*, *mean*, and *standard deviation*, re-

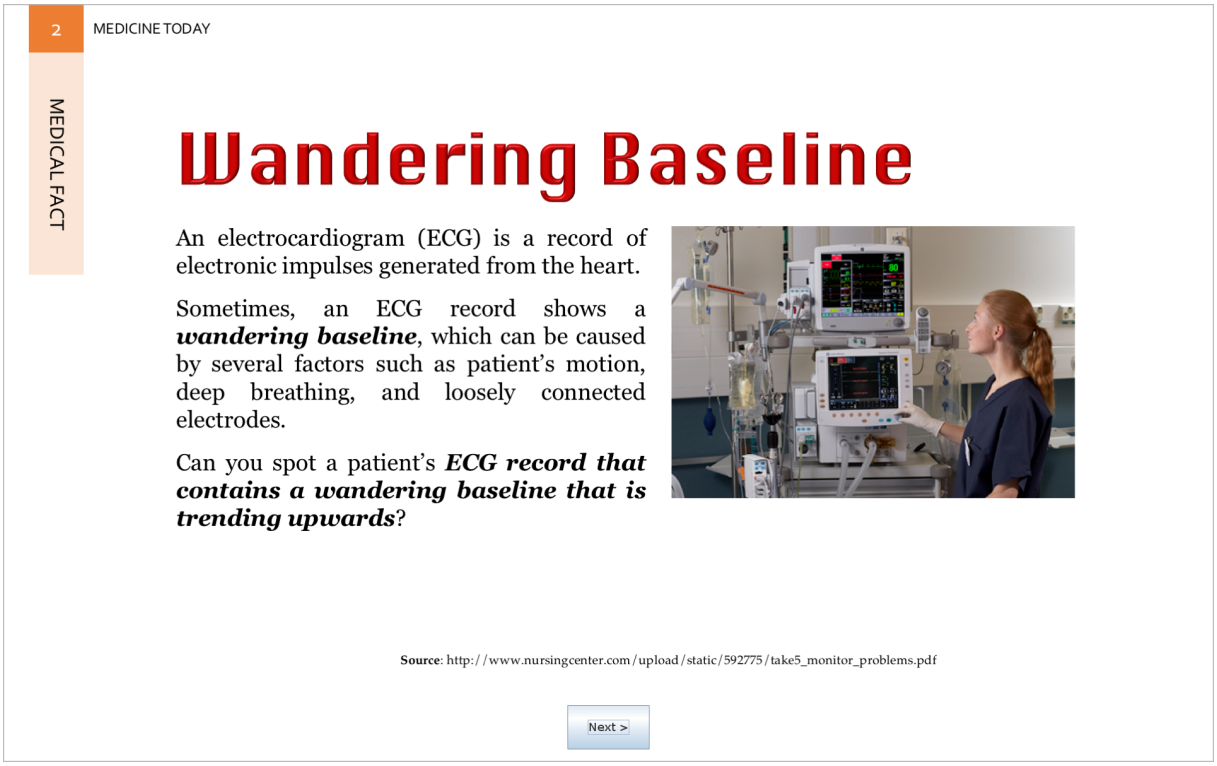


Figure 3: *An information screen in the form of a magazine article, which offers the participants with two clues about the time series to be identified: (i) the contextual type is ECG, and (ii) the specific pattern is a wandering baseline trending upwards.*

Figure 4: *A question screen, consisting of a statistical indicator as the third clue and eight time series plots as optional answers.*

(a) stock market (b) weather temperature (c) electricity production

Figure 5: *Three information screens for the contextual types of stock market, weather temperature, and electricity production.*

flecting the different levels of cognitive load that may be required for reasoning with such measures.

In addition, there are 3 training trials, where electricity production are used as the true contextual type to avoid biases in favor of the three other contextual types (ECG, Stock, and Temperature) used in the main trials. A list of all trials, together with their main attributes, is given in the supplementary materials.

Stimuli Design and Data Virtualization. We conducted an extensive search for information about *electrocardiogram (ECG)*, *stock price*, *weather temperature*, and *electricity production*. We identified various interesting patterns that can be described in the form of a magazine or newspaper article. The source of information is acknowledged on each information screen. For each pattern, we then selected a matching segment of time series data from a data repository in the same application context. We randomly assigned one of the three statistical measures (min/max, mean, stdev), and computed the statistical measure for the data segment. Once we determined the correct answer (TPS, 111) for a trial, we created seven distractors. All such distractors were created based on real-world data, though we had to modify some of the data in order to meet the definition of each distractor. For example, for a distractor (tPs), we first obtained a suitable time series segment in the contextual type specified as a distraction. If the pattern did not quite match the pattern clue, we manipulated the data to feature a matching pattern. We then computed the statistical indicator for the assigned statistical measure. If the measure was too close to the measure of the correct answer, we scaled the data to ensure sufficient difference.

In order to focus the participants on the main visual characteristics of the stimuli, it is necessary to eliminate the potential confounding effects due to non-visual clues (e.g., axis labels) or additional variables (e.g., the number of data points in a plot). We first standardized all time series to 209 data points each, which are displayed with two *virtual axes*. The *virtual time axis* is divided into four sections, and the five grid lines are labeled as $t0$, $t52$, $t104$, $t156$, and $t208$. The *virtual value axis* ranges from 0 to 1000, and is divided into 10 sections, with labels $0, 100, \ldots, 1000$.

We used ECG time series from Physio.net [GAG*00], where typical intervals between two data points are of 0.004 seconds. We plotted segments of data of 2.496 seconds. By aggregating every three original data points, this yields 209 data points per segment. We mapped the original value range from [-5, 5] to [0, 1,000].

The stock market datasets used in this study were from Yahoo Finance [Yah16]. The surface temperature datasets were from the Environmental Protection Agency Average Daily Temperature Archive maintained by the University of Dayton [Knu16]. The electricity production datasets were from ISO New England [ISO16]. For these datasets, the time spans are all set to 4 years per plot. Each data point is an aggregation of one week data. The 209 data points represent $4\times52 + 1$ weeks (1461 days including a leap day).

Stock market prices are in the unit of \$0.01 (USD). When the maximum value of a stock exceeds 1000, we normalized its values by scaling down to the [0, 1000] range. When the values vary within \$0.10 (USD) within a 4-year period, we scaled up the values of the dataset to ensure the range is of at least 10 cents. For all time series representing surface temperatures, we first standardize them with the unit of Fahrenheit. We mapped the range of [-34°F, 100°F] to [0, 1000], which ensures all the datasets that we have used in this study to have non-negative values on the virtual value axis. The amount of electricity produced is normally measured in megawatt hour (Mwh). As different types of fuel have different electricity yield, we normalize them according to each fuel type.

We informed the participants of the two virtual axes, and asked them not to associate axis labels directly with the expected time intervals and data values in any application context. As they were also informed of the overall time period per plot is 2.496 seconds for ECG and 4 years for the other three contextual types, they were expected to estimate mentally any temporal location (e.g., winter) along the virtual time axis when it was required in some trials.

5. Study Implementation

Apparatus. The experiment was supported by a purposely-written software system, which was implemented in Java. It managed the sequences of trials, displayed the information and question screens for each trial, and collected responses from the participants. All time series plots were generated using Microsoft Excel in a resolution of 400×375 pixels. The experiment took place in a computer laboratory at the Department of Computer Science, University of Oxford. The software ran on computers with 3.7 GB of RAM, 3.30 GHz quad-core Intel core i5-3550 processors. The operating system on these computers was Linux Fedora, a Linux with GNOME version 3.4.2. Each computer had a 24-inch Dell's LCD display

with 1920×1200 resolution and in sRBG color mode. Our software ran in the full-screen mode. We adjusted the displays to the same level of brightness and contrast. Participants interacted with the experiment system through a mouse on each desk. A projector was available in the room for the pre-study presentations.

Procedure. Following two pilot studies for evaluating the study design and testing the software, the experiment was conducted in four sessions. Each session consists of 10-13 participants. We set the limit of participants to 15 for each session in order to provide them with an adequate support. The time taken to complete the experiment was approximately 30-65 minutes, excluding the pre-study presentation (15 minutes).

At the beginning of each session, the experimenter gave a brief introductory presentation to the participants. This was aimed to help the participants familiarize themselves with the software, interaction, and tasks. It included a brief explanation about the time series and the three statistical measures used in the study, descriptions of the contextual types concerned in this study and data virtualization, example screenshots of the software system (but not related to the main trials), and instructions related to the task, two-stage design of the trials, and the whole experiment sequence. The participants were informed that they could take as much time as they needed in each trial.

Following the presentation, copies of an information sheet and a consent form were distributed for the participants to read and sign. Each participant was then given a unique user ID, and completed a demographic form for information such as user ID, gender, age group, occupation, and familiarity with time series. The experiment started with a training session consisting of three trials. It then moved to the 36 main trials. After each trial, there was masking screen (white noise) for 2 seconds before the next trial.

After completing all the trials, each participant completed a subjective questionnaire on the ease of identifying a time series with different types of clues.

Participants. A total of 47 participants took part in the study in return for a £10 Amazon voucher. All participants were recruited from the University of Oxford and related communities. Their academic disciplines include computer science, engineering, mathematics, materials, oceanography, anthropology, applied linguistics, economics, and public policy. One participant did not finish the experiment, and his responses were not used in the analysis. Among the 46 participants who completed the study, there were 30 males and 16 females. There were 29 university students and 13 university staff, together with 4 who stated their occupations as "others". 25 participants were in the 20-29 age group, 11 in the 30-39 age group, 6 in the 40-49 age group, 3 in the 50-59 age group, and 1 in the 60-69 age group. In rating their own knowledge about time series, 3 rated "high knowledgeable", 10 "very familiar", 22 "moderately familiar", 8 "heard of it", and 3 "never heard of it".

6. Results and Analysis

In this section, we first examine the impact of the three different classes of knowledge. We then consider each class individually, examining the potential difference within the class.

Table 3: *Statistics of the participants' selections of different optional answers, including **total** of all 46 participants, **percentage** among all 1656 selections, **minimum**, **maximum**, and **average** per participant, and **standard deviation** of the average. The columns are ordered according to the numbers of selections.*

	tps 000	Tps 100	tpS 001	tPs 010	TpS 101	tPS 011	TPs 110	TPS 111
total	13	18	28	41	76	171	178	1131
percent	0.8%	1.1%	1.7%	2.5%	4.6%	10.3%	10.7%	68.3%
min	0	0	0	0	0	0	1	8
max	3	2	6	3	7	10	7	33
average	0.28	0.39	0.61	0.89	1.65	3.72	3.87	24.59
stdev	0.66	0.65	1.31	0.97	1.54	2.53	1.60	4.91

Three Classes of Knowledge. With 36 trials each with 8 optional answers, 46 participants made a total of 1656 selections. As discussed in Sections 3 and 4, the 8 optional answers include only 1 correct answer. Hence the chance due to random clicking is 12.5%. Table 3 summarizes the statistics of these selections in relation to the 8 types of optional answers. The percentage of correct answers is 68.3%, which is well-above the chance. This indicates that in 68.3% of the cases, participants managed to make correct inference based on all three pieces of knowledge. Only in 0.8% of the cases, all three pieces of knowledge failed to guide a participant towards the correct answers.

From Table 3, we can also observe that the numbers of selections correlate with the amount of knowledge that guided participants towards the correct answers. The successful uses of all three pieces of knowledge (111) amount to 68.3%, those with two pieces of knowledge (011, 101, 110) amount to $4.6 \sim 10.7\%$, and those with one piece of knowledge (001, 010, 100) amount to $1.1 \sim 2.5\%$. The failures to use any knowledge (000) amount to the smallest number of selections, i.e., 0.8%. We can also observe that the percentage of each group of incorrect selections (i.e., $000, 001, \ldots, 110$) is below the chance of 12.5%.

A one-way ANOVA (Analysis of variance) was applied to all eight groups of selections (i.e., $000 \sim 111$), resulting in $F(7,360) = 642.78; p < 0.01$, confirming the significant differences among these groups. Further analysis using two-tail paired t-test shows that among the 28 pairwise group comparisons, 23 pairs exhibit significant differences ($p < 0.01$). The difference between Groups 000 (tps) and 001 (tpS) is of marginally significant ($p = 0.05$). Only four pairs, 000 (tps) and 100 (Tps), 100 (Tps) and 001 (tpS), 001 (tpS) and 010 (tPs), and 011 (tPS) and 110 (TPs), show insignificant difference.

While all three categories of knowledge have positive impact in guiding participants towards the correct answers, they appear to have different levels of impact. Figure 6 summarizes the role of three categories of knowledge in cases of partially-correct cases, i.e., excluding 000 (tps) and 111 (TPS). Here we consider that a partially-correct selection reflects positively on the piece of knowledge that was used to narrow down the options, and negatively on a piece of knowledge that did not remove uncertainty successfully. For example, the selection group 110 (TPs) indicates that participants have used the knowledge of contextual types and specific pat-

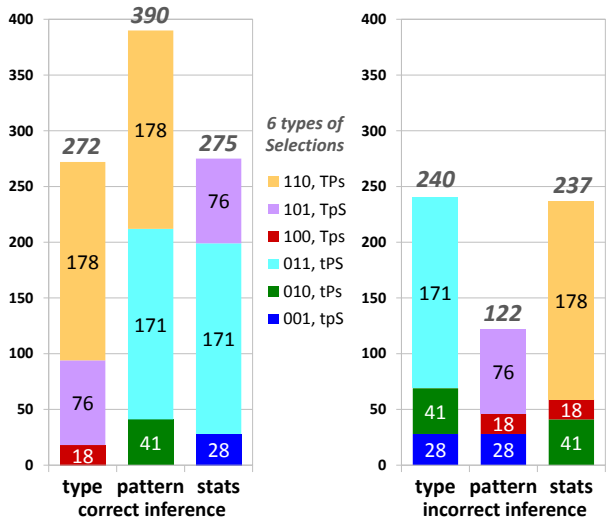


Figure 6: *The numbers of partially-correct selections. Left: the numbers of cases where a form of knowledge was used to reduce the uncertainty successfully. Right: the numbers of cases where a form of knowledge failed to reduce the uncertainty.*

terns correctly, but failed to use the knowledge of statistical measures successfully. Hence, partially-correct groups 100, 110, and 101 reflect the knowledge of contextual types positively, while 001, 010, and 011 reflect such knowledge negatively. From Figure 6, we can observe that the knowledge of contextual types and statistical measures seems to have very similar level of impact, while the knowledge of specific patterns contributes more towards correct inference, and less towards incorrect inference. Note that among these three categories of knowledge, the ability to recognize patterns in data is the most difficult for machines to acquire.

Contextual Type: ECG vs. Stock vs. Temperature. We first made a comparison among the three groups of samples, i.e., all participants' responses to all ECG trials, and all responses to Stock trials, and all responses to Temperature trials. Table 4 summarizes the results. We can observe that the contextual type of "Temperature" resulted in higher accuracy (ACC). The t-tests with Bonferroni correction also confirm that in terms of accuracy the differences in Temperature vs. ECG and Temperature vs. Stock are significant. Meanwhile the difference in ECG vs. Stock is insignificant. On the other hand, ECG demanded the lowest response time (RT). The t-tests with Bonferroni correction confirm that ECG vs. Stock and ECG vs. Temperature have significant variation, while Stock vs. Temperature is insignificant.

Because Mauchly's Test of Sphericity indicates that the assumption of sphericity has been violated, Huynh-Feldt Corrections were applied to the ANOVA results. In order to be sure about the above global observation among the three groups, we also analyzed the samples of the three contextual types under each of the six conditions representing all combinations of a pattern class and a statistical measure, i.e., [simple, complex]×[min/max, mean, stdev]. The ANOVA results are shown in Table 5, where "(HF)" stands for Huynh-Feldt Correction. From this table, we can observe that the variations in response time are present in all conditions, where

Table 4: *Global analysis of three contextual types.*

	ECG	Stock	Temperature
mean ACC	0.643	0.639	0.766
stdev ACC	0.480	0.481	0.424
Sphericity	$p = 0.0466$ (violation)		
ANOVA	$F(2,550) = 16.3, p < 0.01$ (Huynh-Feldt $\varepsilon = 0.99$)		
mean RT	30.75s	35.20s	37.68s
stdev RT	21.91	31.34	28.59
Sphericity	$p < 0.01$ (violation)		
ANOVA	$F(2,550) = 16.3, p < 0.01$ (Huynh-Feldt $\varepsilon = 0.94$)		

Table 5: *Condition-specific analysis of three contextual types.*

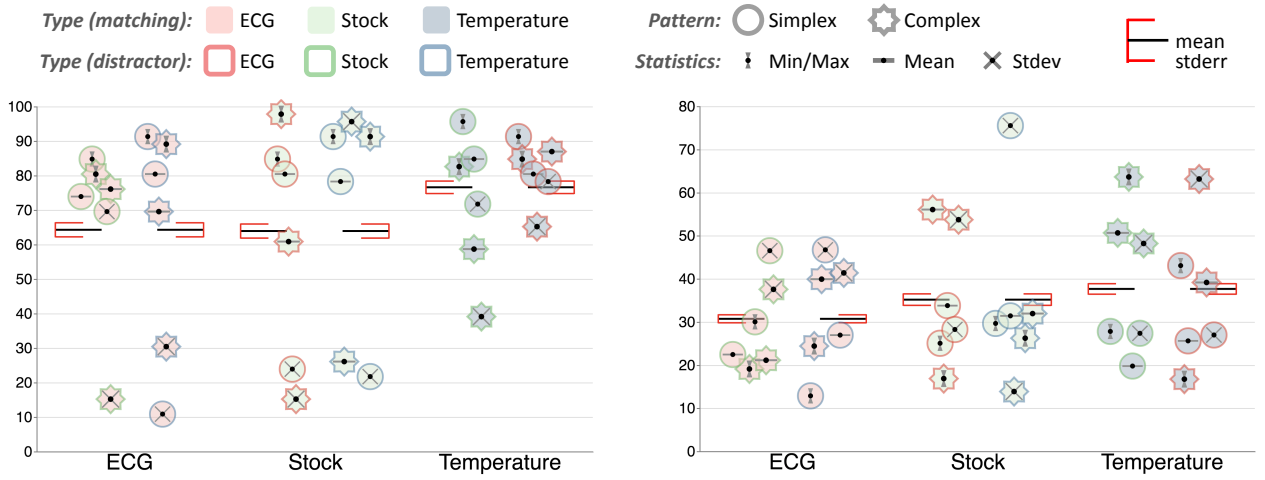
	Pattern	Stats	Sphericity	$F(2,90)$	p value
ACC	simple	min/max	yes	1.595	insignificant
ACC	simple	mean	yes	0.439	insignificant
ACC	simple	stdev	yes	37.97	< 0.01
ACC	complex	min/max	no	4.526	insignificant
ACC	complex	mean	yes	12.70	< 0.01
ACC	complex	stdev	no	22.06	< 0.01 (HF)
RT	simple	min/max	no	18.81	< 0.01 (HF)
RT	simple	mean	no	9.091	< 0.01 (HF)
RT	simple	stdev	no	19.25	< 0.01 (HF)
RT	complex	min/max	no	16.31	< 0.01 (HF)
RT	complex	mean	yes	9.953	< 0.01
RT	complex	stdev	yes	18.33	< 0.01

the variations in accuracy are only certain in 3 (out of 6) conditions. It suggests that in terms of effectiveness differences among the three contextual types become more apparent when one encounters a complex patterns or a complex statistical measure.

In addition, we can make some empirical observations using Figure 7, where each glyph represents all samples under a condition [Type (matching), Type (distractor), Pattern, Statistics]. The average value, in terms of either accuracy (%) or response time (in seconds), is used as the y-coordinate of the glyph. The x-coordinate is coarsely determined by the groups shown on the x-axis, but it is randomly shifted left or right to minimize the overlapping with other glyphs in the same group. From Figure 7(a), we can observe that the mean value of group "Temperature" is higher than the means of the other two groups. It is also interesting to see that among the 8 glyphs below the 50% line, 5 have complex patterns, and 7 have "stdev" indicators. Meanwhile, in Figure 7(b), there is an outlier [Stock, Temperature, complex, stdev] that has an average response time below 15 seconds. while most glyphs with low response times are associated with either a "min/max" or "mean" indicator.

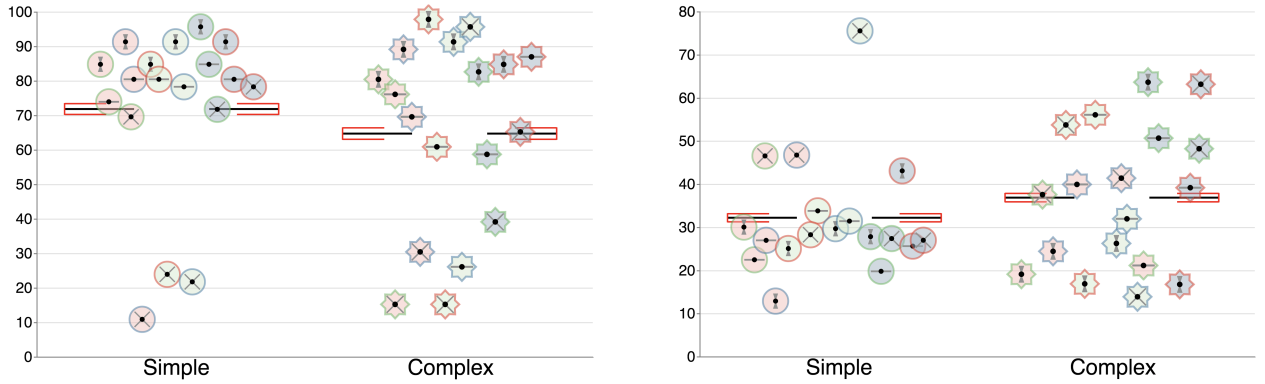
Specific Pattern: Simple vs. Complex. Similar to our approach to the analysis of contextual types, we first divided the samples to two groups of patterns. This allows us to compare the global statistical indicators of simple and complex patterns. Table 6 summarizes the results. We can observe that the simple patterns resulted in higher accuracy and lower response time. The ANOVA confirmed the significant differences in terms of both accuracy and response time.

Although the notion of sphericity does not apply to the two group analysis, we conducted the condition-specific analysis nonetheless. Table 7 shows the analysis results under all 18 conditions. Out of



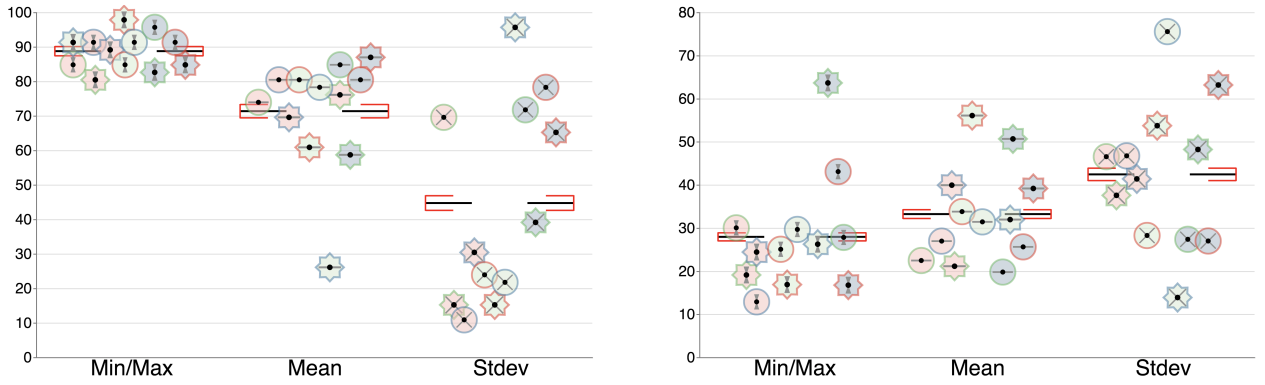
(a) ACC: ECG vs. Stock vs. Temperature

(b) RT: ECG vs. Stock vs. Temperature



(c) ACC: simple vs. complex

(d) RT: simple vs. complex



(e) ACC: min/max vs. mean vs. stdev

(f) RT: min/max vs. mean vs. stdev

Figure 7: *Glyph-based visualization of the 36 trials grouped based on contextual types (top), specific patterns (middle), and statistical measures (bottom). Each glyph encodes the three attributes of a trial. The y-location of each glyph depicts the mean value of the accuracy of a trial (left) or the mean of the response time of a trial (right). The x-location within a group is randomly assigned to minimize overlappings.*

six insignificant results, four are associated with the "min/max" conditions, and two with the "mean" conditions. This suggests that the difference between simple and complex groups is more apparent under the "stdev" conditions.

In addition, we can make some empirical observations using Figure 7. In Figure 7(c), the relative merit of simple patterns is obvious, as most glyphs are above the average lines, and all three glyphs with poor accuracy have "stdev" indicators. Similarly, in Figure 7(d), a large number of glyphs in the simple group are below the average of response time. One glyph featuring "stdev" has added some distor-

Table 6: *Global analysis of two pattern classes.*

	Simple	Complex
mean ACC	0.719	0.647
stdev ACC	0.450	0.478
ANOVA	$F(1,827) = 12.2, p < 0.01$	
mean RT	32.21s	36.88s
stdev RT	26.98	28.22
ANOVA	$F(1,827) = 11.5, p < 0.01$	

Table 7: *Condition-specific analysis of two pattern classes.*

	Type	Stats	$F(1,91)$	p value
ACC	ECG	min/max	0.471	insignificant
ACC	ECG	mean	0.613	insignificant
ACC	ECG	stdev	6.143	0.015
ACC	Stock	min/max	2.617	insignificant
ACC	Stock	mean	30.633	< 0.01
ACC	Stock	stdev	24.580	< 0.01
ACC	Temperature	min/max	4.422	0.038
ACC	Temperature	mean	2.494	insignificant
ACC	Temperature	stdev	12.048	< 0.01
RT	ECG	min/max	0.024	insignificant
RT	ECG	mean	5.292	0.024
RT	ECG	stdev	4.807	0.031
RT	Stock	min/max	8.572	< 0.01
RT	Stock	mean	12.280	< 0.01
RT	Stock	stdev	6.563	0.012
RT	Temperature	min/max	1.036	insignificant
RT	Temperature	mean	37.438	< 0.01
RT	Temperature	stdev	50.871	< 0.01

tion into the average. The glyphs in the complex group are evenly distributed at each side of the average.

Statistical Measure: min/max vs. mean vs. stdev. We first make a global comparison among the three groups of samples, i.e., all participants' responses to "min/max" trials, those to "mean" trials, and those to "stdev" trials. Table 8 summarizes the results. We can easily observe that the contextual type of "stdev" resulted in the lowest accuracy and highest response time. The t-tests with Bonferroni correction also confirm that in terms of both accuracy and response time all pairwise comparison show significant differences. Hence the ordering in terms of effectiveness is clearly "min/max" better than "mean", which is better than "stdev".

Because Mauchly's Test of Sphericity indicates that the assumption of sphericity has been violated, Huynh-Feldt Corrections were applied to the ANOVA results. In order to be sure about the above global observation among three groups, we also analyzed the samples of the three statistical measures under each of the six conditions representing all combinations of a contextual type and a pattern class, i.e., [ECG, Stock, Temperature]×[simple, complex]. The ANOVA results are shown in Table 9, where "(HF)" stands for Huynh-Feldt Correction. From this table, we can observe that all variations are statistically significant.

The interpretation of the above analysis results is consistent with most people's intuition, i.e., given a measure of standard deviation, it would be very hard to imagine a time series that matches such a

Table 8: *Global analysis of three statistical measures.*

	min/max	mean	stdev
mean ACC	0.888	0.714	0.447
stdev ACC	0.316	0.452	0.498
Sphericity	$p < 0.01$ (violation)		
ANOVA	$F(2,550) = 160.2, p < 0.01$ (Huynh-Feldt $\varepsilon = 0.95$)		
mean RT	27.96s	33.23s	42.44s
stdev RT	21.85	23.98	33.82
Sphericity	$p < 0.01$ (violation)		
ANOVA	$F(2,550) = 38.8, p < 0.01$ (Huynh-Feldt $\varepsilon = 0.91$)		

Table 9: *Condition-specific analysis of three statistical measures.*

	Type	Pattern	Sphericity	$F(2,90)$	p value
ACC	ECG	simple	no	32.878	< 0.01 (HF)
ACC	ECG	complex	yes	58.423	< 0.01
ACC	Stock	simple	yes	78.498	< 0.01
ACC	Stock	complex	no	64.539	< 0.01 (HF)
ACC	Temperature	simple	yes	6.601	0.0211
ACC	Temperature	complex	yes	11.577	< 0.01
RT	ECG	simple	no	31.05	< 0.01 (HF)
RT	ECG	complex	no	33.334	< 0.01 (HF)
RT	Stock	simple	no	11.838	< 0.01 (HF)
RT	Stock	complex	yes	19.823	< 0.01
RT	Temperature	simple	yes	10.83	< 0.01
RT	Temperature	complex	no	4.405	< 0.01 (HF)

clue. Meanwhile it is relatively easier to imagine a time series with a given "min/max" or "mean" value. On the other hand, imagine a time series with a complex pattern based on a "mean" value is harder than based on a "min" or "max" value. Therefore, to most human observers, the order of the three clues in terms of effectiveness is "min/max", "mean", and "stdev" (from better to worse).

We can make further observations from Figure 7. Figure 7(e) convincingly depicts the differences among the three statistical measures in terms of accuracy. The patterns in Figure 7(f) are slightly complex for response time, while the ordering of the three measures is apparent. There is a high-RT outlier [Temperature, Stock, complex, min/max] in the "min/max" group, and a low-RT outlier [Stock, Temperature, complex, stdev] in the "stdev" group.

7. Information-Theoretic Measures

Recall the cost-benefit metric in Section 3. We can derive the *Benefit* measurement from the average *accuracy* in Section 6, and approximate the *Cost* measurement using the average *response time*. One must note that any application of the cost-benefit metric to real-world applications is much more complicated as it is necessary to obtain the probability mass functions (PMFs) for the underlying data alphabets and pattern alphabets. In this empirical study, such a PMF is controlled. Each of the three soft models, $\mathbf{M}_T$, $\mathbf{M}_P$, and $\mathbf{M}_S$, is exposed to an alphabet of 8 letters each time. When the 8 options are presented to the participants in each trial, the default assumption is $p = 0.125$ for all 8 letters. The *pretended entropy* is thereby 3 bits. Since there is only one correct answer, the truth PMF has $p = 1$ for one letter, and $p = 0$ for the other seven letters. Hence the *actual entropy* is 0 bits.

To avoid singularity conditions in computing the Kullback-Leibler divergence, it is necessary to make $p = 1 - \varepsilon$ $(0 < \varepsilon < 1)$ for the correct answer, and $p = \varepsilon/7$ for the others. The calculation is unfortunately sensitive to the value of ε. To remove this sensitivity, we first set *the maximum distortion* due to random guess. It is always helpful to set the maximum distortion about twice as much as the maximum compression. In our case, we set $\varepsilon = 0.0063$. Its implication becomes more apparent below.

Let us consider a model that randomly selects an option as an answer. Because $\varepsilon = 0.0063$, the *actual entropy* becomes 0.073 bits. The *Alphabet Compression* in Eq. 1 is thus $3 - 0.073 = 2.927$ bits for this random model. When we apply a reverse mapping from the answer to the truth PMF, there is one in eight chances to be correct. Hence the PMF of the reconstructed alphabet has $p = 0.125$ for all letters. The Kullback-Leibler divergence between the reconstructed PMF and truth PMF results in 5.854 bits of *Potential Distortion*, which is twice as much as the *Alphabet Compression*. The *Benefit* is thus $2.927 - 5.854 = -2.927$ bits for random guess.

We can now compute the *Benefit* for the model that combines $\mathbf{M}_T$, $\mathbf{M}_P$, and $\mathbf{M}_S$. The percentage values in Table 3 naturally become the reconstructed PMF with the correct answer ($p = 1 - \varepsilon$) aligned with TPS ($q = 68.3\%$). The *Potential Distortion* is 1.593 bits. The *Benefit* is thus $2.927 - 1.593 = 1.334$ bits. Similarly, we can derive the *Potential Distortion* of $\mathbf{M}_T$, $\mathbf{M}_P$, and $\mathbf{M}_S$ as 0.508, 0.197, and 0.499 bits respectively. Since their *Alphabet Compression* is 0.945 bit (for $\varepsilon = 0.0063$), the *Benefit* values are 0.437, 0.748, and 0.446 bits respectively. In other words, participants' soft knowledge for recognizing patterns brings the most benefit. For details of the calculations, please refer to the formula in [CG16] and the spreadsheet in the supplementary materials.

8. Conclusions

In this paper, we have reported an empirical study to examine the impact of humans' soft knowledge in reducing potential distortion in data analysis and visualization processes. The participants' successful utilization of all three classes of knowledge together amounts to 68.3% of all responses. This provides the hypothesis in [CG16] with an overwhelming support. Further analysis showed that the participants' knowledge about patterns to be matched with the clues given by textual descriptions performed better than the knowledge about contextual types and statistical measures. Here the definition of "better" assumes that the three classes of knowledge can be used to narrow the probability distribution of a data alphabet equally. Interestingly, from a machine-centric perspective, the knowledge about statistical measures can be defined mathematically, and a reconstruction model can be programmed, for example, by brute-force computation and elimination. The knowledge about contextual types can be obtained by computing a probability distribution of different letters in an alphabet, if one can have access to a huge amount of real-world data defined on the alphabet. However, pattern recognition remains to be a challenge in automated machine intelligence. In comparison, before the experiment, the participants were totally unprepared for what patterns to be discovered in these time series plots, but they performed their tasks with some ease.

The findings of this study do not in any way suggest that we should stop developing machine intelligence. On the contrary, the experiment has confirmed that there were significant differences in the participants' performance in relation to the three statistical measures. Clearly, many participants had difficulties with standard deviation, which incurred more distortion than min/max and mean. Since there are many other statistical measures that most people would have difficulties in mapping them back to the data alphabet, it is highly desirable to use machine-centric processes for such reverse mappings. In general, this manipulation experiment complements the observation experiment in [TKC17], providing another piece of evidence to support the approach of visual analytics, where statistics, algorithms, visualization, and interaction are integrated. In visual analytics, humans' soft knowledge can improve the cost-benefit ratio in inference processes by reducing the potential distortion typically caused by rapid alphabet compression using statistics and algorithms [CG16]. When a visualization process is successful, it is not only because of a clever visual design, but also because it enables humans to use their soft knowledge. We hope that more empirical studies in the future will investigate the contributions of human knowledge in visualization.

References

[ALBR16] ALBO Y., LANIR J., BAK P., RAFAELI S.: Off the radar: Comparative evaluation of radial visualization solutions for composite indicators. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 569–578. 2

[ARH12] AIGNER W., RIND A., HOFFMANN S.: Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions. *Computer Graphics Forum 31*, 3pt2 (2012), 995–1004. 2

[BARM*12] BORGO R., ABDUL-RAHMAN A., MOHAMED F., GRANT P. W., REPPA I., FLORIDI L., CHEN M.: An empirical study on using visual embellishments in visualization. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2759–2768. 2

[BBIF12] BOUKHELIFA N., BEZERIANOS A., ISENBERG T., FEKETE J. D.: Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2769–2778. 2

[BBK*16] BORKIN M. A., BYLINSKII Z., KIM N. W., BAINBRIDGE C. M., YEH C. S., BORKIN D., PFISTER H., OLIVA A.: Beyond memorability: Visualization recognition and recall. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 519–528. 2

[BBL12] BOYANDIN I., BERTINI E., LALANNE D.: A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. *Computer Graphics Forum 31*, 3pt2 (2012), 1005–1014. 2

[BDJ14] BORGO R., DEARDEN J., JONES M. W.: Order of magnitude markers: An empirical study on large magnitude number detection. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 2261–2270. 2

[BDM*17] BEECHAM R., DYKES J., MEULEMANS W., SLINGSBY A., TURKAY C., WOOD J.: Map LineUps: Effects of spatial structure on graphical inference. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 391–400. 2

[BEDF16] BOY J., EVEILLARD L., DETIENNE F., FEKETE J. D.: Suggested interactivity: Seeking perceived affordances for information visualization. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 639–648. 2

[BI12] BEZERIANOS A., ISENBERG P.: Perception of visual variables on tiled wall-sized displays for information visualization applications. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2516–2525. 2

[BNRS13] BRANDES U., NICK B., ROCKSTROH B., STEFFEN A.: Gestaltlines. *Computer Graphics Forum 32*, 3pt2 (2013), 171–180. 2

[BPC*10] BORGO R., PROCTOR K., CHEN M., JAENICKE H., MURRAY T., THORNTON I. M.: Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *IEEE Trans. Visualization & Computer Graphics 16*, 6 (2010), 963–972. 2

[BVB*13] BORKIN M. A., VO A. A., BYLINSKII Z., ISOLA P., SUNKAVALLI S., OLIVA A., PFISTER H.: What makes a visualization memorable? *IEEE Trans. Visualization & Computer Graphics 19*, 12 (2013), 2306–2315. 2

[BW14] BAE J., WATSON B.: Reinforcing visual grouping cues to communicate complex informational structure. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 1973–1982. 2

[CAB*16] CHUNG D. H. S., ARCHAMBAULT D., BORGO R., EDWARDS D. J., LARAMEE R. S., CHEN M.: How ordered is it? On the perceptual orderability of visual channels. *Computer Graphics Forum 35*, 3 (2016), 131–140. 2

[CBH*06] CHEN M., BOTCHEN R. P., HASHIM R. R., WEISKOPF D., ERTL T., THORNTON I. M.: Visual signatures in video visualization. *IEEE Trans. Visualization & Computer Graphics 12*, 5 (2006), 1093–1100. 2

[CDF14] CHEVALIER F., DRAGICEVIC P., FRANCONERI S.: The not-so-staggering effect of staggered animated transitions on visual tracking. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (Dec 2014), 2241–2250. 2

[CG14] CORRELL M., GLEICHER M.: Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (Dec 2014), 2142–2151. 2

[CG16] CHEN M., GOLAN A.: What may visualization processes optimize? *IEEE Trans. Visualization & Computer Graphics 22*, 12 (2016), 2619–2632. 1, 2, 3, 11

[Coh95] COHEN P.: *Empirical Methods for Artificial Intelligence*. The MIT Press, 1995. 2

[DBD17] DIMARA E., BEZERIANOS A., DRAGICEVIC P.: The attraction effect in information visualization. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 471–480. 2

[DLW*17] DASGUPTA A., LEE J. Y., WILSON R., LAFRANCE R. A., CRAMER N., COOK K., PAYNE S.: Familiarity Vs Trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 271–280. 2

[EMdSP*15] ETEMADPOUR R., MOTTA R., D. S. PAIVA J. G., MINGHIM R., DE OLIVEIRA M. C. F., LINSEN L.: Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Trans. Visualization & Computer Graphics 21*, 1 (2015), 81–94. 2

[FHSW13] FINK M., HAUNERT J. H., SPOERHASE J., WOLFF A.: Selecting the aspect ratio of a scatter plot based on its Delaunay triangulation. *IEEE Trans. Visualization & Computer Graphics 19*, 12 (2013), 2326–2335. 2

[FIB*14] FUCHS J., ISENBERG P., BEZERIANOS A., FISCHER F., BERTINI E.: The influence of contour on similarity perception of star glyphs. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 2251–2260. 2

[GAG*00] GOLDBERGER A. L., AMARAL L. A., GLASS L., HAUSDORFF J. M., IVANOV P. C., MARK R. G., MIETUS J. E., MOODY G. B., PENG C.-K., STANLEY H. E.: Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation 101*, 23 (2000), e215–e220. 6

[GBFM16] GSCHWANDTNER T., BÖGL M., FEDERICO P., MIKSCH S.: Visual encodings of temporal uncertainty: A comparative user study. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 539–548. 2

[GCNF13] GLEICHER M., CORRELL M., NOTHELFER C., FRANCONERI S.: Perception of average value in multiclass scatterplots. *IEEE Trans. Visualization & Computer Graphics 19*, 12 (2013), 2316–2325. 2

[GHL15] GUO H., HUANG J., LAIDLAW D. H.: Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE Trans. Visualization & Computer Graphics 21*, 10 (2015), 1173–1186. 2

[GR15] GRIFFIN A. L., ROBINSON A. C.: Comparing color and leader line highlighting strategies in coordinated view geovisualizations. *IEEE Trans. Visualization & Computer Graphics 21*, 3 (2015), 339–349. 2

[GSL14] GRAMAZIO C. C., SCHLOSS K. B., LAIDLAW D. H.: The relation between visualization size, grouping, and user performance. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 1953–1962. 2

[HFMC12] HOFMANN H., FOLLETT L., MAJUMDER M., COOK D.: Graphical tests for power comparison of competing designs. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2441–2448. 2

[HKF16] HAROZ S., KOSARA R., FRANCONERI S. L.: The connected scatterplot for presenting paired time series. *IEEE Trans. Visualization & Computer Graphics 22*, 9 (2016), 2174–2186. 2

[HKH*12] HÖFERLIN M., KURZHALS K., HÖFERLIN B., HEIDEMANN G., WEISKOPF D.: Evaluation of fast-forward video visualization. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2095–2103. 2

[HW12] HAROZ S., WHITNEY D.: How capacity limits of attention influence information visualization effectiveness. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2402–2410. 2

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using Weber's law. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 1943–1952. 2

[ISO16] ISO NEW ENGLAND: Daily generation by fuel type, accessed in 2016. URL: https://www.iso-ne.com/. 6

[JH13] JAKOBSEN M. R., HORNBÃ£K K.: Interactive visualizations on large and small displays: The interrelation of display size, information space, and scale. *IEEE Trans. Visualization & Computer Graphics 19*, 12 (2013), 2336–2345. 2

[JH16] JANSEN Y., HORNBÃ£K K.: A psychophysical investigation of size as a physical variable. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 479–488. 2

[JHKH13] JAKOBSEN M. R., HAILE Y. S., KNUDSEN S., HORNBÃ£K K.: Information visualization and proxemics: Design opportunities and empirical findings. *IEEE Trans. Visualization & Computer Graphics 19*, 12 (2013), 2386–2395. 2

[KARC15] KANJANABOSE R., ABDUL-RAHMAN A., CHEN M.: A multi-task comparative study on scatter plots and parallel coordinates plots. *Computer Graphics Forum 34*, 3 (2015), 261–270. 2

[KDX*12] KIM S. H., DONG Z., XIAN H., UPATISING B., YI J. S.: Does an eye tracker tell the truth about visualizations?: Findings while investigating visualizations for decision making. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2421–2430. 2

[KHW13] KURZHALS K., HÖFERLIN M., WEISKOPF D.: Evaluation of attention-guiding video visualization. *Computer Graphics Forum 32*, 3pt1 (2013), 51–60. 2

[KMLM16] KWON O. H., MUELDER C., LEE K., MA K. L.: A study of layout, rendering, and interaction methods for immersive graph visualization. *IEEE Trans. Visualization & Computer Graphics 22*, 7 (2016), 1802–1815. 2

[Knu16] KNUTH D.: University of Dayton - Environmental Protection Agency Average Daily Temperature Archive, accessed in 2016. URL: http://academic.udayton.edu/kissock/http/Weather/default.htm. 6

[KZZM12] KUANG X., ZHANG H., ZHAO S., MCGUFFIN M.: Tracing tuples across dimensions: A comparison of scatterplots and parallel coordinate plots. *Computer Graphics Forum 31*, 3pt4 (2012), 1365–1374. 2

[LARC16] LICCARDI I., ABDUL-RAHMAN A., CHEN M.: I know where you live: Inferring details of people's lives by visualizing publicly shared location data. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2016), pp. 1–12. 2

[LDA12] LIVINGSTON M. A., DECKER J. W., AI Z.: Evaluation of multivariate visualization on a multivariate task. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2114–2121. 2

[LDM*01] LAIDLAW D. H., DAVIDSON J. S., MILLER T. S., DA SILVA M., KIRBY R. M., WARREN W. H., TARR M.: Quantitative comparative evaluation of 2D vector field visualization methods. In *Proc. IEEE Visualization* (2001), pp. 143–150. 2

[LFK*13] LIN S., FORTUNA J., KULKARNI C., STONE M., HEER J.: Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum 32*, 3pt4 (2013), 401–410. 2

[LMVW10] LI J., MARTENS J.-B., VAN WIJK J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (2010), 13–30. 2

[MDF12] MICALLEF L., DRAGICEVIC P., FEKETE J. D.: Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2536–2545. 2

[MK15] MITTELSTÄDT S., KEIM D. A.: Efficient contrast effect compensation with personalized perception models. *Computer Graphics Forum 34*, 3 (2015), 211–220. 2

[MPWG12] MARRIOTT K., PURCHASE H., WYBROW M., GONCU C.: Memorability of visual features in network diagrams. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2477–2485. 2

[MRO*12] MACEACHREN A. M., ROTH R. E., O'BRIEN J., LI B., SWINGLEY D., GAHEGAN M.: Visual semiotics & uncertainty visualization: An empirical study. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2496–2505. 2

[NBW14] NETZEL R., BURCH M., WEISKOPF D.: Comparative eye tracking study on node-link visualizations of trajectories. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 2221–2230. 2

[NHB*17] NETZEL R., HLAWATSCH M., BURCH M., BALAKRISHNAN S., SCHMAUDER H., WEISKOPF D.: An evaluation of visual search support in maps. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 421–430. 2

[OPH*16] OTTLEY A., PECK E. M., HARRISON L. T., AFERGAN D., ZIEMKIEWICZ C., TAYLOR H. A., HAN P. K. J., CHANG R.: Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 529–538. 2

[PQMCR17] PADILLA L., QUINAN P. S., MEYER M., CREEM-REGEHR S. H.: Evaluating the impact of binning 2D scalar fields. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 431–440. 2

[RB10] RENSINK R. A., BALDRIDGE G.: The perception of correlation in scatterplots. *Computer Graphics Forum 29*, 3 (2010), 1203–1210. 2

[RKSB13] RAGAN E. D., KOPPER R., SCHUCHARDT P., BOWMAN D. A.: Studying the effects of stereo, head tracking, and field of regard on a small-scale spatial judgment task. *IEEE Trans. Visualization & Computer Graphics 19*, 5 (2013), 886–896. 2

[SHK15] SKAU D., HARRISON L., KOSARA R.: An evaluation of the impact of visual embellishments in bar charts. *Computer Graphics Forum 34*, 3 (2015), 221–230. 2

[SK16] SKAU D., KOSARA R.: Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum 35*, 3 (2016), 121–130. 2

[SOK*16] STROBELT H., OELKE D., KWON B. C., SCHRECK T., PFISTER H.: Guidelines for effective usage of text highlighting techniques. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 489–498. 2

[SSK16] SAKET B., SCHEIDEGGER C., KOBOUROV S.: Comparing node-link and node-link-group visualizations from an enjoyment perspective. *Computer Graphics Forum 35*, 3 (2016), 41–50. 2

[SSKB14] SAKET B., SIMONETTO P., KOBOUROV S., BÖRNER K.: Node, node-link, and node-link-group diagrams: An evaluation. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 2231–2240. 2

[SSKB15] SAKET B., SCHEIDEGGER C., KOBOUROV S. G., BÃŰRNER K.: Map-based visualizations increase recall accuracy of data. *Computer Graphics Forum 34*, 3 (2015), 441–450. 2

[STM17] SARVGHAD A., TORY M., MAHYAR N.: Visualizing dimension coverage to support exploratory analysis. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 21–30. 2

[TGH12] TALBOT J., GERTH J., HANRAHAN P.: An empirical model of slope ratio comparisons. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2613–2620. 2

[TKC17] TAM G. K. L., KOTHARI V., CHEN M.: An analysis of machine- and human-analytics in classification. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017). 1, 2, 3, 4, 11

[TLM16] TANAHASHI Y., LEAF N., MA K.-L.: A study on designing effective introductory materials for information visualization. *Computer Graphics Forum 35*, 7 (2016), 117–126. 2

[TSA14] TALBOT J., SETLUR V., ANAND A.: Four experiments on the perception of bar charts. *IEEE Trans. Visualization & Computer Graphics 20*, 12 (2014), 2152–2160. 2

[VBC*16] VOLANTE M., BABU S. V., CHATURVEDI H., NEWSOME N., EBRAHIMI E., ROY T., DAILY S. B., FASOLINO T.: Effects of virtual human appearance fidelity on emotion contagion in affective interpersonal simulations. *IEEE Trans. Visualization & Computer Graphics 22*, 4 (2016), 1326–1335. 2

[VTW*12] VANDE MOERE A., TOMITSCH M., WIMMER C., CHRISTOPH B., GRECHENIG T.: Evaluating the effect of style in information visualization. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2739–2748. 2

[WBJ16] WALKER J., BORGO R., JONES M. W.: TimeNotes: A study on effective chart visualization and interaction techniques for time-series data. *IEEE Trans. Visualization & Computer Graphics 22*, 1 (2016), 549–558. 2

[WCA*17] WU Y., CAO N., ARCHAMBAULT D., SHEN Q., QU H., CUI W.: Evaluation of graph sampling: A visualization perspective. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 401–410. 2

[WPHC16] WUN T., PAYNE J., HURON S., CARPENDALE S.: Comparing bar chart authoring with Microsoft Excel and tangible tiles. *Computer Graphics Forum 35*, 3 (2016), 111–120. 2

[XRP*12] XU K., ROONEY C., PASSMORE P., HAM D. H., NGUYEN P. H.: A user study on curved edges in graph visualization. *IEEE Trans. Visualization & Computer Graphics 18*, 12 (2012), 2449–2456. 2

[Yah16] YAHOO: Yahoo! Finance, accessed in 2016. URL: https://finance.yahoo.com. 6

[YDGM17] YANG Y., DWYER T., GOODWIN S., MARRIOTT K.: Many-to-many geographically-embedded flow visualisation: An evaluation. *IEEE Trans. Visualization & Computer Graphics 23*, 1 (2017), 411–420. 2

[ZOC*13] ZIEMKIEWICZ C., OTTLEY A., CROUSER R. J., YAUILLA A. R., SU S. L., RIBARSKY W., CHANG R.: How visualization layout relates to locus of control and other personality factors. *IEEE Trans. Visualization & Computer Graphics 19*, 7 (2013), 1109–1121. 2

[ZWM13] ZHENG L., WU Y., MA K. L.: Perceptually-based depth-ordering enhancement for direct volume rendering. *IEEE Trans. Visualization & Computer Graphics 19*, 3 (2013), 446–459. 2